GLOBUSONLINE OVERVIEW

Tanya Levshina

WHAT IS GLOBUSONLINE (GO)?

- GO is a Software-as-a-Service facility that provides file transfer functionality.
- It will do a third-party transfer on your behalf:
  - Performs transfers of files
  - Retries in case of failures
  - Optimizes gridftp parameters for transfer
    - Concurrency
    - Threading
    - Pipelining
  - Provides CLI, Rest API, WEB UI
  - Handles certificate via MyProxy servers or uses your proxy certificate
  - Provides user support

WHY DO WE(OSG) CARE?

- Big VOs, with dedicated storage and staff, have their own data management services (FTS, PhedEx, SAM-Grid, …)
- A VO that relies on OSG public storage encounters common problems:
  1. What SE is available for my VO?
  2. How much space can I use?
  3. What is the most efficient way to transfer significant amount of data (GBs - TBs) from on site to another?
  4. What is the best way to handle errors and retires?
  5. How to monitor the transfer progress?
- GO may help us with #3-5

GO GLOSSARY

- **EndPoints**
  - **Logical Name**
  - **Transfer servers**
  - **Activation Status**
  - **MyProxy Server**
- **Transfer Task**
  - **Task ID**
  - **Status**
  - **Start Time**
  - **End Time**
  - **Deadline**
  - **Event Log**
  - **Sync level**
  - **Subtasks List**
    - **Subtask ID**
    - **Status**
    - **Start Time**
    - **End Time**
    - **Event Log**

HOW CAN I USE IT?

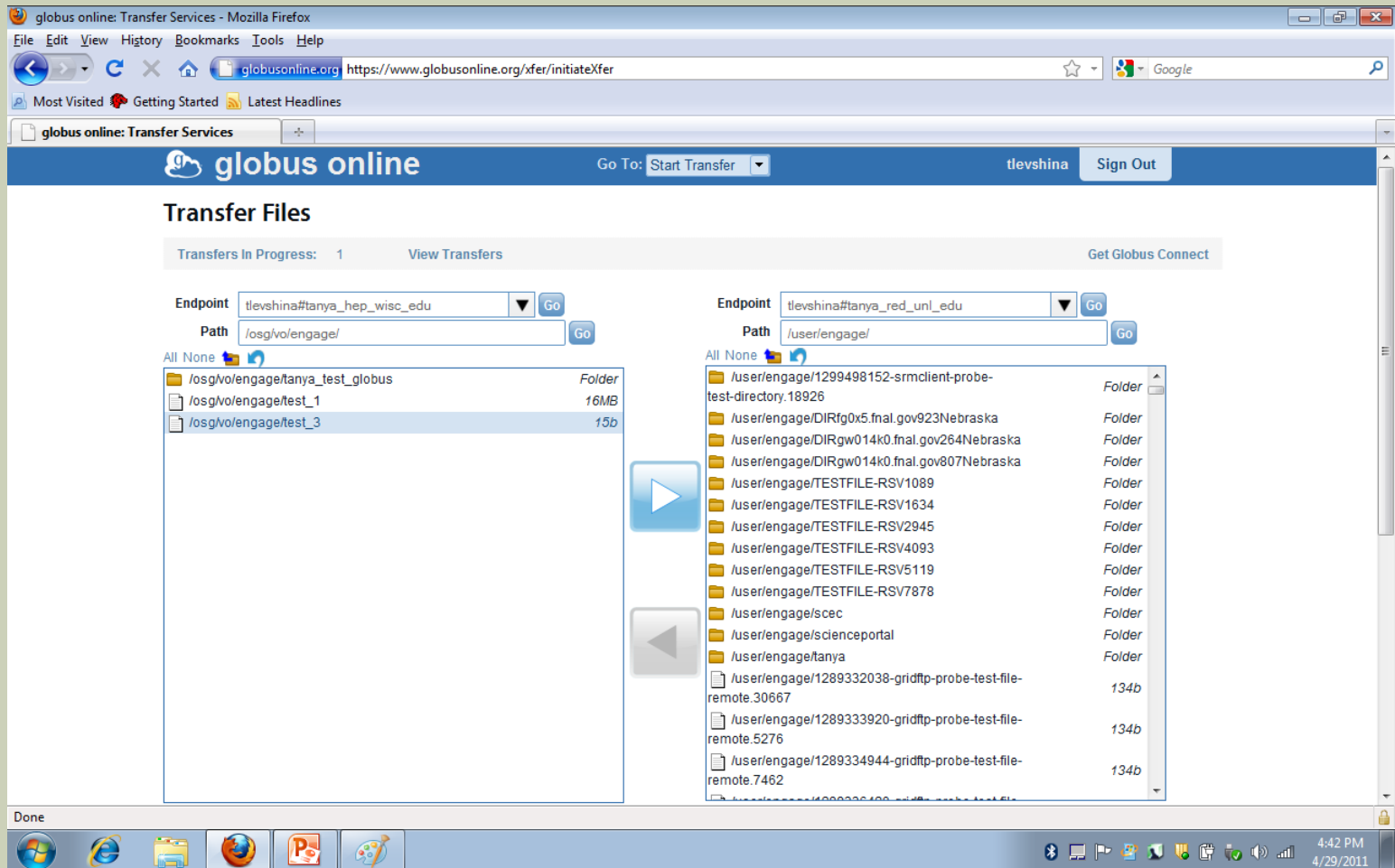- Sunny day scenario:
  - Storage Admin:
    - Creates public endpoints that work for your VO
  - User
    - Prepares the list of files you want to transfer from A to B
    - Gets proxy certificate and activate endpoint with your proxy or use myProxy server
    - Initializes the transfer using UI, CLI or Rest API
    - Monitors the progress of transfer
    - Gets notification when job succeeded

TRANSFER WEB UI

WEB MONITORING (I)

WEB MONITORING (II)

GO CLI

- **Requirements (available in OSG Client):**
  - grid-proxy-init
  - gssish
- **Create/modify/activate/list endpoint**
  ```
  $ gsissh -t tlevshina@cli.globusonline.org endpoint-list
  go#ep1                    20:40:53
  gw_hdfs                   162:35:31
  hep_wisc_edu_1               89:53:38
  tanya_hep_wisc_edu           89:53:38
  tanya_red_unl_edu            190:11:12

  $ gsissh -t tlevshina@cli.globusonline.org endpoint-list -v tanya_red_unl_edu
  Name           : tanya_red_unl_edu
  Host(s)        : gsiftp://red-gridftp1.unl.edu:2811
  Subject(s)     :
  MyProxy Server   : n/a
  Credential Status : ACTIVE
  Credential Expires: 2011-05-07 20:17:26Z Credential Subject: /DC=org/DC=doegrids/OU=People/CN=Tanya Levshina
  508821/CN=614089312/CN=383514127
  ```
- **List directories and files**
  ```
  gsissh -t tlevshina@cli.globusonline.org ls tanya_red_unl_edu:/user/engage/tanya_test_ftp_1
  ```
- **File copy (scp)**
  ```
  $ gsissh -t tlevshina@cli.globusonline.org scp tanya_red_unl_edu:/user/engage/tanya_test_ftp_1 tanya_hep_wisc_edu:/osg/vo/engage/test_1
  Task ID: c8c776ba-72ad-11e0-aa60-1231380e306c
  [XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX] 1/1 0.00 mbps
  ```
  - **Get status and transfer details**
  ```
  $ gsissh -t tlevshina@cli.globusonline.org status c8c776ba-72ad-11e0-aa60-1231380e306c
  Task ID    : c8c776ba-72ad-11e0-aa60-1231380e306c
  Request Time: 2011-04-29 22:12:34Z
  Command    : scp tanya_red_unl_edu:/user/engage/tanya_test_ftp_1 tanya_hep_wisc_edu:/osg/vo/engage/test_1
  Status     : SUCCEEDED
  ```
  - **Access to event logs**
- **File Transfer**
  - Read EOF or CNTRL-D terminated list (can specify sync level)
  - Can set expiration time

REST API

- Python and Java API is available, not too difficult use Rest API with PERL
- Good documentation
- Easy to integrate with existing code
  - Integrated and tested with DES workflow during two days of integration-fest at Fermilab
- Access to more detailed information about transfer status and event
- Adequate support
- Requirements:
  - Python 2.6 or higher

TESTS RESULTS

- DES
  - 31,000 files
  - 184 GB
  - 27% improvement in transfer time
  - DES workflow requires simple transfer verification (file length comparison after transfer). So, total transfer and verification time was similar to the time of original workflow.

- LIGO emulation
  - 5,000 files
  - 16MB each (78GB)
  - GridFTP servers on top of HDFS
  - Average transfer time between Nebraska and Wisconsin is slightly better then the transfer time shown by LIGO tests with multiple srm-copy commands with concurrency 5

ISSUES

- Number of retries per file is not supported. Only expiration time can be specified.
- SRM is not supported, have to explicitly list all grdiftp servers you want to use. GO gidftp-servers load balancing could interfere with SRM load balancing.
- Do not distinguish between "FATAL" and intermittent errors during file transfer: continue retries in case of FATAL error.
  - Several OSG SEs uses HDFS that doesn't allow "seek" operation, so any attempt to retransmit data in case of failure with offset >0 fails.
- GO does not provide interfaces to do verification (get stats) of a list of files.
- Some features are missing from API (for example: activation with certificate).
- Doesn't provide space accounting.

HOW CAN IT BE USED IN THE OSG ENVIROMENT?

- In order to use OSG public storage we need to:
  - Discover what OSG SEs support a specific VO (discovery tools can do it).
  - Use pigeon tools to verify that SE is working for this VO.
  - Use GO CLI or Rest API to dynamically create endpoints.
  - Prepare list of files or directories for transfer.
  - Use GO CLI or Rest API to initiate transfer.
  - Use GO CLI, Rest API or WEB UI to monitor and check status.
- Still missing:
  - Accurate account of space utilization and availability.

IF YOU WANT TO TRY IT NOW...

- **Register with GO**
- **Generate proxy certificate**
  voms-proxy-init
- **Create your own endpoints for the sites that are working for you (e.g Nebraska and Wisconsin)**
  endpoint-create
- **Configure them with grdiftp servers.**
  endpoint-add, endpoint-activate
  - **Tip: Use discovery tools to find gridftp servers for the site:**
    $ get_gridftp_storage_element_id --vo Engage --show_site_name |grep -i Nebraska
    Nebraska   red-srm1.unl.edu
    $ get_gridftp_url --vo Engage --storage_element_id red-srm1.unl.edu
    gsiftp://red-gridftp1.unl.edu:2811/mnt/hadoop/user/engage/TESTFILE
- **Test:**
  ls
  scp
- **Need help?**
  - support@globusonline.org
  - tlevshin@fnal.gov